Noname manuscript No.
(will be inserted by the editor)

Multimodal Emotion Recognition in Speech-based Interaction Using Facial Expression, Body Gesture and Acoustic Analysis

Loic Kessous · Ginevra Castellano · George Caridakis

Received: date / Accepted: date

Abstract In this paper a study on multimodal automatic emotion recognition during a speech-based interaction is presented. A database was constructed consisting of people pronouncing a sentence in a scenario where they interacted with an agent using speech. Ten people pronounced a sentence corresponding to a command while making 8 different emotional expressions. Gender was equally represented, with speakers of several different native languages including French, German, Greek and Italian. Facial expression, gesture and acoustic analysis of speech were used to extract features relevant to emotion. For the automatic classification of unimodal data, bimodal data and multimodal data, a system based on a Bayesian classifier was used. After performing an automatic classification of each modality, the different modalities were combined using a multimodal approach. Fusion of the modalities at the feature level (before running the classifier) and at the results level (combining results from classifier from each modality) were compared. Fusing the multimodal data resulted in a large increase in the recognition rates in comparison to the unimodal systems: the multimodal approach increased the recognition rate by more than 10% when compared to the most successful unimodal system. Bimodal emotion recognition based on all combinations of the modalities (i.e., 'face-gesture', 'face-speech' and 'gesture-speech') was also investigated. The results show that the best pairing is 'gesture-speech'. Using all three modalities resulted in a 3.3% classification improvement over the best bimodal results.

Keywords Affective body language · Affective speech · Facial Expression · Emotion recognition · Multimodal fusion

L. Kessous
Independent Researcher, 30 chemin du Lancier Marseille, France, 13008 E-mail: loic.kessous@gmail.com

G. Castellano
Department of Computer Science, School of Electronic Engineering and Computer Science, Queen Mary University of London, Mile End Road, London E1 4NS
Telephone: +44 (0)20 7882 3234
E-mail: ginevra@dcs.qmul.ac.uk

G. Caridakis
Image, Video and Multimedia Systems Laboratory School of Electrical and Computer Engineering National Technical University of Athens
E-mail: gcari@image.ece.ntua.gr

1 Introduction

Emotion is generally expressed through several modalities in human-human interaction. In some cases, when one of the modalities is missing, there can be confusion about the meaning and the comprehension of the expressed emotion. For example, defective sound during a video conference can induce confusion in the perception of the speakers' emotion by listeners. This is particularly true when a person expressing an emotion assumes that the visual and audio modalities of the communication will be transferred, but when in fact his/her interlocutor is not receiving all of them.

A fake smile hiding disagreement, for instance, might be misinterpreted if the affective content conveyed by the voice is not received by the interlocutor. In this scenario, in fact, the users are not interacting face-to-face, and multimodal visual cues, although proven effective in the automatic discrimination between posed and spontaneous smiles [1], might not be clearly interpreted.

In the field of Human-Machine Interaction based on automatic speech recognition, recognition of emotion is

challenging. From the human perspective, a system endowed with an emotional intelligence should be capable of creating an affective interaction with users: it must have the ability to perceive, interpret, express and regulate emotions [2]. Under these conditions, interacting with a machine would be more similar to interacting with humans and should be more pleasant. From the machine perspective, recognizing the user's emotional state is one of the main requirements for computers to successfully interact with humans [3]. Identification of expressiveness and emotion would improve the understanding of the meaning conveyed by the communication process and could possibly provide a basis for auto-regulation of the system by differentiating between satisfaction and dissatisfaction of the user.

Many related works in affective computing do not combine different modalities into a single system for the analysis of human emotional behavior: different channels of information (mainly facial expressions and speech) are usually considered independently to each other. Further, there have been relatively few attempts to also consider the integration of information from body movement and gestures. Nevertheless, Sebe et al. [4] and Pantic et al. [5] make the point that an ideal system for automatic analysis and recognition of human affective information should be multimodal, just as the human sensory system is. Moreover, studies from psychology highlight the need to consider the integration of different behavior modalities in human-human communication [6].

In this paper a multimodal approach for the recognition of eight acted emotional states (*Anger*, *Despair*, *Interest*, *Pleasure*, *Sadness*, *Irritation*, *Joy* and *Pride*) is presented. The approach integrates information from facial expressions, body gesture and speech. A model with a Bayesian classifier was trained and tested, using a multimodal corpus with ten subjects, collected during the Third Summer School of the HUMAINE EU-IST project, held in Genova in September 2006.

The main contribution of this study consists of integrating three different modalities for the purpose of emotion recognition. Bimodal emotion recognition based on all combinations of the modalities is also investigated. To date, some efforts have been made to build systems capable of recognizing emotions based on two modalities (e.g., based on the combination of facial expressions and speech data [7] and facial expressions and gesture [8]). However, the use of three modalities is still poorly explored. Karpouzis et al. [9] proposed a multicue approach based on facial, vocal and bodily expressions to model affective states, but the fusion of modalities is modelled at the level of facial expressions and speech data only. The present work goes further and provides a multimodal framework for emotion recognition, in which different classifiers are trained using all three modalities.

A second contribution of this work is the integration of body gesture information in a framework for emotion recognition. Though the body gesture modality has not been investigated in as much depth as the face has, a number of studies have been proposed in which gesture is used to infer emotions (e.g., [10], [11], [12]). Nevertheless, apart from a few exceptions (e.g., [8], [1]), the fusion of body gesture with other modalities remains mostly unexplored.

Another contribution of this work is to use features that convey information about how the emotional expressions vary over time. Moreover, results seem to suggest that traditional statistical features are not as effective in discriminating between emotions as those referring to the timing of the temporal profile of facial, vocal and bodily expressions.

The objective of this paper is the discrimination of different emotional expressions based on facial expressions, body gesture and speech information. Performances of unimodal, bimodal and multimodal systems are compared. It is expected that the fusion of two modalities will increase the recognition rate in comparison with the use of one modality only. An improvement in the performance when the three modalities are simultaneously used is also expected.

Results show that the combination of two modalities increases the performance of the classifier when facial expressions and speech data are fused together, as well as when body gesture and speech information is combined. The combination of facial expressions and body gesture, however, improves the results of the classifier based on facial expressions, but not the classifier based on body gesture. Results also show that the fusion of three modalities allows for the highest recognition rate to be obtained.

In the following Sections, after a short review of the state of the art in emotion recognition, we describe the data collection and the feature extraction process. The proposed approach is presented, first by focusing on the analysis performed for each of the three modalities considered in this work and, secondly, based on the fusion of the modalities. Finally, different strategies for performing the data fusion for bimodal and multimodal emotion recognition are compared.

2 Related Work

Generally speaking, emotion recognition based on acoustic analysis has been investigated with three main types

of databases: acted emotions, natural spontaneous emotions and elicited emotions.

Data obtained in the acted situations contain less ambiguous emotions, because actors express the exact emotions they were instructed to. Of course different actors can understand and interpret an instruction differently. This illustrates the importance of the director, and of a good definition of the instructions. Spontaneous speech can, for example, be collected from call center data [13], or interaction with robots [14]. Because of this, emotion collection is more diversified and, often, in order to perform automatic classification, the data must be mapped onto a limited number of classes. Dividing a corpus into categories is a complex task and a non-pertinent grouping can have direct consequences on the recognition rate. When the emotion is elicited, as for example in a "*Wizard of Oz*" scenario [14], the task of dividing the corpus into classes is probably as complex as doing so for spontaneous speech. This is, firstly, because it is highly dependent on user personality and, secondly, because it depends on the context of interaction during the data collection: the more the context is restricted, the less dispersion will be found in the emotion labelling. For the aforementioned reasons, even if it is evident that emotion research should ideally target natural databases, acted databases are useful because it is easier to determine a correspondence between the collected data and their labels.

The best results are therefore generally obtained with acted emotion databases. Literature on speech (see for example Banse and Scherer [15]) shows that the majority of studies have been conducted with emotional acted speech. Feature sets for acted and spontaneous speech have recently been compared in [16]. Generally, few acted-emotion speech databases have included speakers with several different native languages. More recently, some attempts to collect multimodal data were made: some examples of multimodal databases can be found in [17], [18], [19].

In the area of unimodal emotion recognition, there have been many studies using a variety of different, but single, modalities. Facial expressions [20], [21], [22], vocal features [23] [24] [25], body movements and postures [26], [27], [11], [28], physiological signals [29] have been used as inputs during these attempts, although multimodal emotion recognition is currently gaining ground [7], [30], [31], [32], [33]. Nevertheless, most of the work has considered the integration of information from facial expressions and speech [34], [35] and there have been relatively few attempts to combine information from body movement and gestures in a multimodal framework. Gunes and Piccardi [8], for example, fused facial expressions and body gestures at different levels for bimodal emotion recognition. Further, el Kaliouby and Robinson [36] proposed a vision-based computational model to infer acted mental states from head movements and facial expressions. Additionally many psychological studies have highlighted the need to consider the integration of multiple modalities for a proper inference of emotions [6], [37].

A wide variety of machine learning techniques have been used in emotion recognition approaches [21], [3]. Particularly in the multimodal case, they all employ a large number of audio, visual or physiological features, a fact which usually impedes the training process. Therefore, it is necessary to find a way to reduce the number of used features by choosing only those related to emotion. One possibility in this direction is to use neural networks, since they allow the most relevant features with respect to the output to be defined, usually by observing their weights. An interesting approach in this area is sensitivity analysis conducted by Engelbrecht et al. [38]. Sebe et al. [4] highlight that probabilistic graphical models, such as Hidden Markov Models, Bayesian networks and Dynamic Bayesian networks are very well suited for fusing different sources of information when conducting multimodal emotion recognition and can also handle noisy features and missing values of features by probabilistic inference.

In this work we combine a wrapper feature selection approach and a Bayesian classifier. The former reduces the number of features and the latter was used for unimodal, bimodal and multimodal emotion recognition.

3 Collection of multimodal data

The corpus used in this study was collected during the Third Summer School of the HUMAINE EU-IST project, held in Genova in September 2006. The recording procedure was based on that of the GEMEP corpus [18], a multimodal collection of portrayed emotional expressions. Data on facial expressions, body movement and gestures and speech was simultaneously recorded. The development of a new corpus of emotional expressions presents the disadvantage of not being able to compare results with those reported in the literature. Nevertheless, the aim of this study was to build a framework for emotion recognition based on the integration of multiple modalities. Existing databases primarily contain facial and vocal expressions, while gesture is often not included, or accompanied solely by facial expressions. The need of a corpus with three modalities of expression in an interaction scenario was the main motivation leading to our development of a new collection of emotional expressions.

3.1 Subjects and set-up

Ten participants from the summer school, distributed as evenly as possible concerning their gender, participated in the recordings. Participants represented five different nationalities: French, German, Greek, Israeli and Italian. In terms of the technical set-up, two DV cameras (25 fps) recorded the participants from a frontal view. One camera recorded the participants' body and the other one was focused on the participants' face.

We chose such a setup because the resolution required for the extraction of facial features is much higher than that required for body movement detection or hand gesture tracking. This could only be achieved if one camera zoomed in on the participants' face. We adopted some restrictions concerning the participants' behavior and clothing. Long sleeves were preferred since most hand detection algorithms are based on color tracking. Further, a uniform background was used to make the background subtraction process easier. For the facial features extraction process we considered some prerequisites such as an absence of eyeglasses, beards, and moustaches.

For the voice recordings, we used a direct-to-disk computer-based system. The speech samples were directly recorded on the hard disk of the computer using sound editing software. We used an external sound card connected to the computer by an IEEE 1394 High Speed Serial Bus (also known as FireWire or i.Link). A microphone mounted on the participants' shirt was connected to an HF emitter (wireless system emitter) and the receiver was connected to the sound card using a XLR connector (balanced audio connector for high quality microphones and connections between equipment). The external sound card included a preamplifier (for two XLR inputs) that was used in order to adjust the input gain and to minimize the impact of the signal-to-noise ratio of the recording system. The sampling rate of the recording was 44.1 kHz and the quantization was 16 bit, mono.

3.2 Procedure

Participants were asked to act eight emotional states: *Anger*, *Despair*, *Interest*, *Pleasure*, *Sadness*, *Irritation*, *Joy* and *Pride*. We chose this set of features in order to obtain emotions that are equally distributed in valence-arousal space (see Table 1). During the recording process, one of the authors had the role of director, guiding the participants through the process. Participants were asked to perform specific gestures that exemplify each emotion. The director's role was to instruct the participant on the procedure (number of gesture repetitions, emotion sequence, etc.) and details of each emotion and emotion-specific gesture. For example, for the *Despair* emotion the participant was given a brief description of the emotion (e.g.: *"facing an existential problem without solution, coupled with a refusal to accept the situation"*).

In case the subjects failed to follow the experiment script or misjudged the instructed emotion, the researcher acting as the director provided further clarifications or an illustrative example of an occurrence of such an emotion to the subject.

In case of despair, this scenario was *"You have just learned that your father has been diagnosed with a very advanced cancer. According to the doctors, there is not much hope. You are not able to make peace with this idea and you seek by all means to find a solution which the doctors have not thought of. However, you know well that it is without hope"*. All instructions were provided by taking inspiration from the procedure and scenarios used during the collection of the GEMEP corpus [18]. For selecting the emotion-specific gestures, we borrowed ideas from figure animation research dealing with posturing of a figure [39] and elaborated the gestures shown in Table 1.

Table 1 The acted emotions and emotion-specific gestures.

Emotion	Valence	Arousal	Gesture
Anger	Negative	High	Violent descend of hands
Despair	Negative	High	Leave me alone
Interest	Positive	Low	Raise hands
Pleasure	Positive	Low	Open hands
Sadness	Negative	Low	Smooth falling hands
Irritation	Negative	Low	Smooth go away
Joy	Positive	High	Circular italianate movement
Pride	Positive	High	Close hands towards chest

As in the GEMEP corpus [18], a pseudo-linguistic sentence was pronounced by the participants while they acted out the emotional states. The sentence *"Toko, damato ma gali sa"* was designed in order to fulfil different needs. First, as the different participants had different native languages, using a specific language was not adequate for this study. Using a language that is native for one of the participants would have made the task easier for him than for the others, and using a language native to none of them might have favored one who was more fluent in the language than the others.

We also wanted the sentence to include phonemes that exist in all the languages of all the participants. Also, the words in the sentence are composed of simple diphones (*'ma'* and *'sa'*), two (*'gali'*,*'toko'*) or three

diphones (*'damato'*). In addition, the vowels included (*'o', 'a', 'i'*) are relatively distant in vowel space (for example the vowel triangle), and have a similar pronunciation in all the languages of the participants' group. We suggested to the participants that they communicate a conveyed message for the sentence. *"Toko"* is supposed to be the name of a person (or a robot, a virtual agent or a command system), who the participants are interacting with. For this word, we chose two stop consonants (also known as plosives or stop-plosives) /t/ and /k/ and two identical vowels /o/. This was done in order to allow the study of certain acoustic correlates. Then "damato ma gali sa" is meant to represent a verbal command or request (such as, for example, "*can you open it*". The word "*it*" could correspond to a folder, a file, a box, a door and so on. Each emotion was acted out three times by each participant, resulting in the collection of 240 posed gestures, facial expressions and speech samples.

4 Feature extraction

4.1 Face feature extraction

Initially a face detection algorithm is applied to the image to detect the position and boundaries of the face in the foreground (Figure 1). From the plethora of face detection algorithms [40] we selected the Viola Jones algorithm, which is actually a cascade of boosted classifiers working with haar-like features. Since position is provided by this algorithm, the head roll rotation (around +Z axis) can be estimated according to the line connecting the pupils of the two eyes; the face region is then rotated so that this line is parallel with Y axis. Afterwards, the rectangular face boundary region is segmented into coarse facial feature candidate areas, containing the features whose boundaries need to be extracted, according to anthropometric measurements [41], focusing on the left eye/eyebrow, right eye/eyebrow, nose and mouth. This approach minimizes the search area for facial feature boundaries into a small proportion of the entire image, thus speeding up the feature extraction process. For every facial feature, a multi-cue approach is adopted, generating a number of masks which are produced by a number of algorithms [42] performing well under different lighting conditions and resolutions. Feature masks generated for each facial feature are fused together to produce the final mask for that feature. The mask fusion process uses anthropometric criteria [41] to perform validation and weight assignment on each intermediate mask; all feature weighted masks are then fused to produce a final mask along with confidence level estimation.

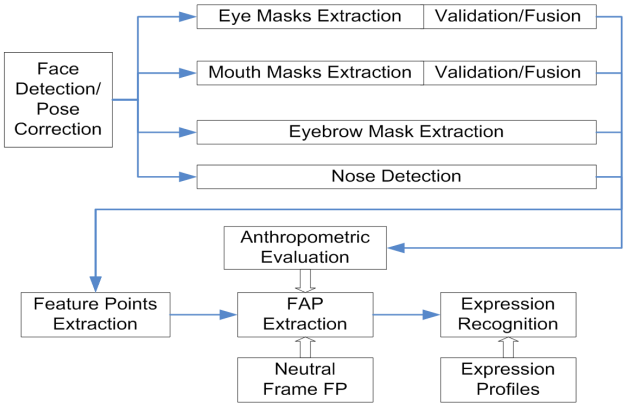

Fig. 1 Face feature extraction

We chose to work with MPEG 4 FAPs (Facial Animation Parameters) and not Action Units (AUs), since our procedure essentially locates and tracks points in the facial area and the former are explicitly defined to measure the deformation of these feature points. Measurement of FAPs requires the availability of a frame where the participants' expression is found to be neutral. This frame is called the *neutral frame* and is manually selected from the video sequences to be analyzed or interactively provided to the system in the initial phase of interaction. The final feature masks were used to extract 19 Feature Points (FPs) [43]; FPs obtained from each frame were compared to FPs obtained from the neutral frame in order to estimate facial deformations and produce the FAPs. Confidence levels on FAP estimation were derived from the equivalent feature point confidence levels. The FAPs were used along with their confidence levels to provide the facial expression estimation.

In accordance with the other modalities, facial features needed to be processed so as to obtain one vector of values per sentence. FAPs originally correspond to every frame in the sentence. Our method for imprinting the temporal evolution of the FAP values was to calculate the set of statistical features over time of these values and their derivatives. The whole process was inspired by the equivalent process performed for the acoustic features.

The most common problems, especially encountered in low quality input images or situations of illumination changes and complex and/or dynamic backgrounds, include connection with other feature boundaries, and mask dislocation due to noise. In some cases, masks may have completely missed their goal and provide a completely invalid result. Outliers such as illumination changes and compression artifacts cannot be predicted and so individual masks have to be re-evaluated and combined on each new frame. This calculation process

takes place on a per frame basis and the mask fusion technique shields the feature extraction algorithm against lighting condition changes and dynamic background situations. Of course, all the steps rely on the Viola Jones face detector which has been proven to be very robust and adaptive. Overall, our algorithm performs well under large variations of facial image quality, color and resolution.

In an attempt to validate the proposed facial feature extraction algorithm, 250 frames (randomly selected) were manually annotated from two human observers and the group agreement metric was calculated [42]. This metric was the Williams's Index (WI), which actually divides the average number of agreements (inverse disagreements) between the computer (observer 0) and human observers by the average number of agreements between human observers. At a value of 0, the computer mask is infinitely far from the observer mask. When WI is larger than 1, the computer generated mask disagrees less with the observers than the observers disagree with each other. The WI for each facial region was 0.838, 0.875, 0.780, 1.034 and 1.013 for left eye, right eye, mouth, left eyebrow and right eyebrow respectively.

4.2 Body feature extraction

The tracking of the body and hands of the participants was conducted using the EyesWeb platform [44]. Starting from the silhouette and the *blobs* representing the hands of the participants, five main expressive motion cues were extracted using the EyesWeb Expressive Gesture Processing Library [45]: Quantity of Motion (QoM) and Contraction Index (CI) of the body, Velocity (VEL), Acceleration (ACC) and Fluidity (FL) of the hand's barycenter.

The Quantity of Motion (QoM) is a measure of the amount of detected motion, computed with a technique based on silhouette motion images (SMIs). These are images carrying information about variations in the silhouette shape and position in the last few frames.

$$SMI[t] = \sum_{i=0}^{n} Silhouette[t-i]\} - Silhouette[t] \qquad (1)$$

The SMI at frame t is generated by adding together the silhouettes extracted in the previous n frames and then subtracting the silhouette at frame t. The resulting image contains the variations that occurred in the previous frames.

QoM is computed as the area (i.e., number of pixels) of a SMI, normalized in order to obtain a value usually ranging from 0 to 1. It can be considered as an overall measure of the amount of detected motion, involving velocity and force.

$$QoM = Area(SMI[t,n])/Area(Silhouette[t]) \qquad (2)$$

The Contraction Index (CI) is a measure, ranging from 0 to 1, of the degree of contraction and expansion of the body. CI can be calculated using a technique related to the bounding region, i.e., the minimum rectangle surrounding the body: the algorithm compares the area covered by this rectangle with the area currently covered by the silhouette.

Velocity (VEL) and acceleration (ACC) are related to the trajectory followed by the hand's barycenter in a 2D plane. Fluidity gives a measure of the uniformity of motion, so that fluidity is considered maximum when, in the movement between two specific points of the space, the acceleration is equal to zero. It is computed as the Directness Index [45] of the trajectory followed by the velocity of hand's barycenter in the 2D plane.

Data was normalized according to the behavior shown by each participant, by considering the maximum and the minimum values of each motion cue in each subject, in order to compare data from all the participants.

Automatic extraction allows for temporal series of the selected motion cues over time to be obtained, depending on the video frame rate. Based on the model proposed in [28], for each temporal profile of the motion cues, a subset of features describing the dynamics of the cues over time was extracted (see list below):

- **Initial and Final Slope**: slope of the line joining the first value and the first relative extremum, slope of the line joining the last value and the last relative extremum.
- **Initial (Final) Slope of the Main Peak**: slope of the line joining the absolute maximum and the preceding (following) minimum.
- **Maximum, Mean, Mean / Max, Mean / Following Max**: the maximum and mean values and their ratio, ratio between the two first biggest values.
- **Maximum / Main Peak Duration, Main Peak Duration / Duration**: ratio between the maximum and the main peak duration, ratio between the peak containing the absolute maximum and the total gesture duration.
- **Centroid of Energy, Distance between Max and Centroid**: location of the barycenter of energy, distance between the maximum and the barycenter of energy.
- **Shift Index of the Maximum, Symmetry Index**: position of the maximum with respect to the center of the curve, symmetry of the curve relative to the maximum value position.

– **Number of Maxima, Number of Maxima preceding the Main One**: number of relative maxima, number of relative maxima preceding the absolute one.

Automatic extraction of the selected features was conducted using software modules developed in EyesWeb. This process was done for each motion cue of all the videos of the corpus, so that each gesture is characterized by a subset of 80 motion features.

4.3 Speech feature extraction

The set of features that we used for speech includes features based on intensity, pitch, MFCC (Mel Frequency Cepstral Coefficient), Bark spectral bands, voiced segment characteristics and pause length. The full set contains 377 features. The features from the intensity contour and the pitch contour were extracted using a set of 32 statistical features. This set of features was applied both to the pitch and intensity contour and to their derivatives. Normalization was not applied before feature extraction. In particular, we didn't perform user or gender normalization for pitch contour, as it is often used in order to remove the difference between registers. We considered the following 32 features: maximum, mean and minimum values, sample mode (most frequently occurring value), interquartile range (difference between the 75th and 25th percentiles), kurtosis, the third central sample moment, first (slope) and second coefficients of linear regression, first, second and third coefficients of quadratic regression, percentiles at 2.5%, 25%, 50%, 75%, and 97.5%, skewness, standard deviation and variance. Thus, we have 64 features based on the pitch contour and 64 features based on the intensity contour.

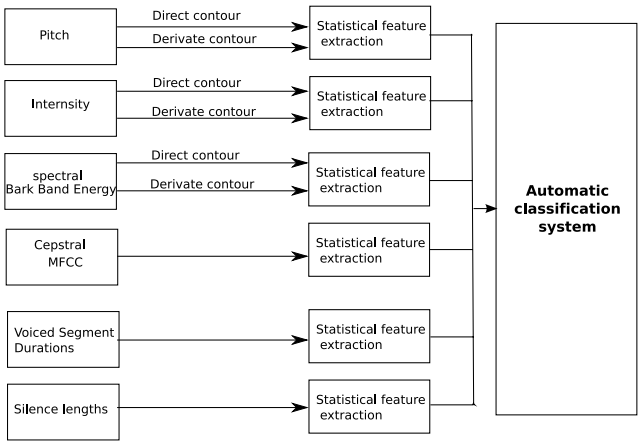

Fig. 2 Speech feature extraction

This feature set was originally used for inspecting a contour, for example, a pitch contour or a loudness contour, but these features are also meaningful for inspecting evolution over time or spectral axis. Indeed, we also extracted similar features on the Bark spectral bands as done in [46]. Further, we extracted 13 MFCCs using time averaging on time windows, as well as features derived from pitch values and lengths of voiced segments using a set of 35 features applied to both of them. Finally, we extracted features based on pause (or silence) length and non-pauses lengths (35 each). This process is summarized in Figure 2.

5 A framework for emotion recognition using multiple modalities

In order to compare the results of the unimodal, bimodal and the multimodal systems, we used a common approach based on a Bayesian classifier (*BayesNet*) provided by the software Weka, a free toolbox containing a collection of machine learning algorithms for data mining tasks [47].

The first algorithm used is a Bayesian network. The estimator algorithm for finding the conditional probability tables of the Bayesian network is *SimpleEstimator*. SimpleEstimator is used for estimating the conditional probability tables of the Bayesian network once the structure has been learned. It estimates probabilities directly from data. The Alpha parameter was set to the value 0.5. Alpha is used for estimating the probability tables and can be interpreted as the initial count on each value. A *K2* learning algorithm was used as the search algorithm for searching network structures. This Bayesian network learning algorithm uses a hill-climbing algorithm restricted by an order on the variables from Cooper and Herskovits [48]. The initial network used for structure learning is a Naïve Bayes Network, that is, a network with a connection from the classifier node to every other node.

In Figure 3, we describe an overview of the framework. As shown in the left side of the figure, a separate Bayesian classifier was used for each modality (face, gestures, speech). All sets of data were normalized using the *normalize* function provided by the software Weka. Feature discretization based on Kononenko's MDL (Minimum Description Length) criterion [49] was conducted to reduce the learning complexity. A wrapper approach to feature subset selection (which allows an evaluation of the attribute sets by using a learning scheme) was used in order to reduce the number of inputs to the

classifiers and find the features that maximize the performance of the classifier.

This algorithm, called *WrapperSubsetEval*, evaluates attribute sets by using a learning scheme. Cross-validation is used to estimate the accuracy of the learning scheme for a set of attributes. The number of folds to use when estimating subset accuracy is set to 5 and the seed to use for randomly generating splits is 1. The threshold to repeat if the standard deviation of mean exceeds is set to the value 0.01. A best-first search method in the forward direction was used. Further, in all the systems, the corpus was trained and tested using a 10-fold cross-validation method.

In K-fold cross-validation, the original data set (the database) is partitioned into K subsets. Of the K subsets, a single subset is retained as the validation data for testing the model, and the remaining K-1 subsets are used as training data. The cross-validation process is repeated K times (*the folds*), with each of the K subsets used exactly once as the validation data. The K results from the folds can then be averaged (or otherwise combined) to produce a single estimation. The advantage of this method over repeated random division into subsets is that observations are used for both training and validation, and each observation is used for validation exactly once. 10-fold cross validation is commonly used, as in our study. For a more detailed description of the K-fold cross-validation and other methods, see [50].

To fuse facial expressions, gestures and speech information, two different approaches were implemented (Figure 3): feature-level fusion, where a single classifier with features of two (for bimodal emotion recognition) or three modalities (for multimodal emotion recognition) is used; and decision-level fusion, where a separate classifier is used for each modality and the outputs are combined a posteriori. In the second approach, the output was computed by combining the posterior probabilities of the unimodal systems. In the second case, we used the same classifier for all the modalities (and the same feature selection process). The classifier was the same as that used for feature-level fusion. We conducted experiments using two different approaches for decision-level fusion. The first approach for decision-level fusion consisted of selecting the emotion that received the highest probability in the three modalities (best probability approach). The second approach for decision-level fusion (majority voting plus best probability) consisted of selecting the emotion that corresponded to the majority voting from the three modalities; if a majority was not possible to define (for example when each unimodal system outputs a different emotion), the emotion that received the highest probability in the three modalities was selected.

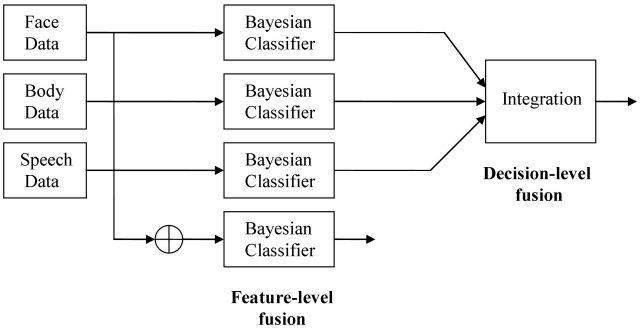

Fig. 3 Overview of the framework

6 Results

6.1 Unimodal Emotion Recognition

Emotion recognition from facial expressions : Table 2 shows the confusion matrix for the emotion recognition system based on facial expressions. The overall performance (percentage of instances correctly classified) of this classifier was **48.3**%. The most recognized emotions were *Anger* (56.67%), *Irritation*, *Joy* and *Pleasure* (53.33%). *Pride* is misclassified as *Pleasure* (20%), while *Sadness* is misclassified as *Irritation* (20%), an emotion in the same valence-arousal quadrant. After the feature selection process, 26 features remain (see Table 5).

Table 2 Confusion matrix of the emotion recognition system based on facial expressions.

a	b	c	d	e	f	g	h	Emotion
56.67	3.33	3.33	10	6.67	10	6.67	3.33	a *Anger*
10	**40**	13.33	10	0	13.33	3.33	10	b *Despair*
6.67	3.33	**50**	6.67	6.67	10	16.67	0	c *Interest*
10	6.67	10	**53.33**	3.33	6.67	3.33	6.67	d *Irritation*
3.33	0	13.33	16.67	**53.33**	10	0	3.33	e *Joy*
6.67	13.33	6.67	0	6.67	**53.33**	13.33	0	f *Pleasure*
6.67	3.33	16.67	6.67	13.33	20	**33.33**	0	g *Pride*
3.33	6.67	3.33	20	0	13.33	6.67	**46.67**	h *Sadness*

Emotion recognition from body gestures : Table 3 shows the performance of the emotion recognition system based on body gesture. The overall performance of this classifier was **67.1**%. *Anger* and *Pride* are recognized with very high accuracy (80% and 96.67% respectively). *Sadness* was partially misclassified as *Pride* (36.67%). After the feature selection process, 18 features remain (see Table 5).

Emotion recognition from speech : Table 4 displays the confusion matrix for the emotion recognition system based on speech. The overall performance of this classifier was **57.1**%. *Anger* and *Sadness* are classified with high accuracy (93.33% and 76.67% respectively). *Despair* obtained a very low recognition rate and was

Table 3 Confusion matrix of the emotion recognition system based on gestures.

a	b	c	d	e	f	g	h	Emotion
80	10	0	3.33	0	0	6.67	0	a *Anger*
3.33	**56.67**	6.67	0	0	0	26.67	6.67	b *Despair*
3.33	0	**56.67**	0	6.67	6.67	26.67	0	c *Interest*
0	10	0	**63.33**	0	0	26.67	0	d *Irritation*
0	10	0	6.67	**60**	0	23.33	0	e *Joy*
0	6.67	3.33	0	0	**66.67**	23.33	0	f *Pleasure*
0	0	0	3.33	0	0	**96.67**	0	g *Pride*
0	3.33	0	3.33	0	0	36.67	**56.67**	h *Sadness*

mainly confused with *Pleasure* (23.33%). After the feature selection process, 18 features remain (see Table 5).

Table 4 Confusion matrix of the emotion recognition system based on speech.

a	b	c	d	e	f	g	h	Emotion
93.33	0	3.33	3.33	0	0	0	0	a *Anger*
10	**23.33**	16.67	6.67	3.33	23.33	3.33	13.33	b *Despair*
6.67	0	**60**	10	0	16.67	3.33	3.33	c *Interest*
13.33	3.33	10	**50**	3.33	3.33	13.33	3.33	d *Irritation*
20	0	10	13.33	**43.33**	10	3.33	0	e *Joy*
3.33	6.67	6.67	6.67	0	**53.33**	6.67	16.67	f *Pleasure*
3.33	10	3.33	13.33	0	13.33	**56.67**	0	g *Pride*
0	6.67	3.33	10	0	3.33	0	**76.67**	h *Sadness*

Table 5 Description of the 10 first selected features for unimodal classifications based on body gesture, speech and facial expressions

Maximum-QoM	gesture	Quantity of Motion	Maximum
Mean-QoM	gesture	Quantity of Motion	Mean
MeanMax-QoM	gesture	Quantity of Motion	Ratio between mean and maximum
MaxFollMax-QoM	gesture	Quantity of Motion	Ratio between maximum and following absolute max
InSlope-CI	gesture	Contraction Index	Initial slope
NPeaks-CI	gesture	Contraction Index	Number of peaks
MeanMax-CI	gesture	Contraction Index	Ratio between mean and maximum
MaxCentroid-CI	gesture	Contraction Index	Distance between maximum and centroid of energy
PeakDurGestDur-CI	gesture	Contraction Index	Ratio between main peak duration and gesture duration
FinalSlope-VEL	gesture	Velocity	Final slope
Pitch-min	speech	pitch	Minimum
Pitch-p2c1	speech	pitch	1st coef. of quad. regression
Pitch-q875	speech	pitch	Quantile 0.875
pv-skew	speech	voiced part	Skewness
Pause-p2c1	speech	pause	1st coef. of quad. regression
Segment-max	speech	Segment duration	Maximum
Segment-mean	speech	Segment duration	Mean
Segment-kurt	speech	Segment duration	Kurtosis
mean-mfcc-06	speech	MFCC	mean of 6th coefficient
mean-mfcc-10	speech	MFCC	mean of 10th coefficient
open-jaw-p2c1	Face	Vertical jaw displacement	1st coef. of quad. regression
open-jaw-q975	Face	Vertical jaw displacement	Quantile 0.975
open-jaw-q90	Face	Vertical jaw displacement	Quantile 0.90
lower-top-midlip-range	Face	Vertical top middle inner lip displacement	Range
raise-bottom-midlip-extdt	Face	Vertical bottom middle inner lip displacement	derivative
widening-mouth-range	Face	Horizontal displacement of inner lip corners	Range
widening-mouth-std	Face	Horizontal displacement of inner lip corners	Standard Deviation
widening-mouth-kurt	Face	Horizontal displacement of inner lip corners	Kurtosis
widening-mouth-range2	Face	Horizontal displacement of inner lip corners	Interquartile Range
close-left-eye-max	Face	Vertical displacement of left eyelids	maximum

The fact that the misclassification does not concern the same classes in the different modalities is encouraging. In this way, we hope that, when using the three modalities, a misclassification in one class will be attenuated in the two others. Our subjective observations lead us to suspect that the misclassifications and confusions may correspond to similarities observed in the expression of the concerned emotions in each modality.

6.2 Feature-level fusion

Table 6 displays the confusion matrix of the multimodal emotion recognition system. The overall performance of this classifier was **78.3**%, which is much higher than the performance obtained by our most successful unimodal system, that based on gestures. The diagonal components reveal that all the emotions, apart from *Despair*, can be recognized with over 70% accuracy. *Anger* was the emotion recognized with highest accuracy, as was the case in all the unimodal systems.

Table 6 Confusion matrix of the multimodal emotion recognition system.

a	b	c	d	e	f	g	h	Emotion
90	0	0	0	10	0	0	0	a *Anger*
0	**53.33**	3.33	16.67	6.67	0	10	10	b *Despair*
6.67	0	**73.33**	13.33	0	3.33	3.33	0	c *Interest*
0	6.67	0	**76.67**	6.67	3.33	0	6.67	d *Irritation*
0	0	0	0	**93.33**	0	6.67	0	e *Joy*
0	3.33	3.33	13.33	3.33	**70**	6.67	0	f *Pleasure*
3.33	3.33	0	3.33	0	0	**86.67**	3.33	g *Pride*
0	0	0	16.67	0	0	0	**83.33**	h *Sadness*

After feature selection, 17 features remain in the final feature set (see Table 7). 5 features are from the gesture modality, 2 features are from the face modality and 9 features are from the speech (acoustic) modality. The number of features remaining for each modality should not be considered as an indication of the contribution of the modality. One possible interpretation is to regard the number of features as an indication of non-redundancy (in the two other modalities) of information. It is, of course, a simplification that should be taken into account regarding other aspects of the results.

As far as body gesture is concerned, the features conserved after the feature selection process are the following: the symmetry of the temporal profile of the Quantity of Motion, the ratio between the mean and maximum value of the Contraction Index and the final slope of the temporal profile of Velocity, Acceleration and Fluidity.

For speech, the selected features include the mean of the absolute deviation of intensity, two coefficients of quadratic regression of the pitch contour, the range between first and last quartile (IQR) of the pitch contour along the sentence, the second coefficient of quadratic regression of the pitch contour at the beginning of the sentence, the maximum pause time in the sentence and the time of the maximum length of the voiced segments. Two more features conserved after the feature selection process are related to Bark spectral band energies. The first is a statistical feature related to the time evolution of one of the spectral bands; the second models the evolution over time of the kurtosis of the spectrum divided into bark spectral bands.

For facial expressions, the features conserved after feature selection are: the range over time of the vertical (downwards direction) displacement of the jaw and

the kurtosis over time of the mouth widening. The feature defining the mouth widening is an intuitive fusion of two animation parameters which correspond to the horizontal displacement of the left and the right inner lip corner and the left and the right displacement motion direction, respectively.

It is interesting to notice that the set of features conserved after the feature selection process includes features that come from all the different modalities. While there are only two remaining facial features, these features make an important difference. Adding facial expression information, in fact, allows for an improvement of 3.3% to be obtained in comparison with the performance achieved by the classifier based on the bimodal pairing of body gesture and speech, as shown in Section 6.4.

Table 7 Selected features for multimodal classification using feature-level fusion.

Symmetry-QoM	gesture	Quantity of Motion	Symmetry
MeanMax-CI	gesture	Contraction Index	Ratio between mean and maximum
FinalSlope-VEL	gesture	Velocity	Final slope
FinalSlope-ACC	gesture	Acceleration	Final slope
FinalSlope-FL	gesture	Fluency	Final slope
Intens-mad0	speech	Intensity	MAD
Pitch-p1c2	speech	Pitch	first order regression
Pitch-p2c1	speech	Pitch	second order
Pitch-range2	speech	Pitch	interquartile range
pv-p1c2	speech	Voiced segment	first order regression
Pause-tmax	speech	Pause	time of maximum
Segmt-tmax	speech	segment	time of maximum
BarkTL-sgxt	speech	Bark spectral bands	time line
BarkSL-kurt	speech	Bark spectral bands	spectral line kurtosis
open-jaw-range	face	Vertical jaw displacement	range
widening-mouth-kurt	face	Horizontal displacement of inner lip corners	kurtosis

6.3 Decision-level fusion

The approach based on decision-level fusion obtained lower recognition rates than that based on feature-level fusion. The performance of the classifier was **74.6**%, both for the *best probability* and for the *majority voting plus best probability* approaches. Table 8 shows the performance of the system with decision-level integration using the best probability approach. *Anger* was again the emotion recognized with highest accuracy, but the recognition rate of the majority of emotions decreases with respect to the recognition rate achieved while performing integration at the feature-level.

Table 8 Decision level integration with best probability approach.

a	b	c	d	e	f	g	h	Emotion
96.67	0	0	0	0	0	3.33	0	a *Anger*
13.33	**53.33**	6.67	0	0	3.33	13.33	10	b *Despair*
3.33	0	**60**	3.33	10	13.33	6.67	3.33	c *Interest*
13.33	6.67	6.67	**60**	0	3.33	0	10	d *Irritation*
0	0	10	3.33	**86.67**	0	0	0	e *Joy*
6.67	3.33	0	0	0	**80**	6.67	3.33	f *Pleasure*
3.33	0	6.67	0	0	10	**80**	0	g *Pride*
3.33	3.33	0	10	0	3.33	0	**80**	h *Sadness*

6.4 Bimodal classification

Using feature-level fusion, we also performed bimodal classification. All combinations of modalities were investigated using the same methods for feature selection and classification as for unimodal and multimodal emotion recognition. When using speech and face modalities, **62.5%** of instances were correctly classified. This result is better than the result obtained for speech only (57.1%). It suggests that facial expressions provide extra emotional information in addition to the speech. This result is in accordance with early studies showing that the face provides complementary information [51] and also with more recent work using natural databases [52], [9]. Improvements obtained for the fusion of speech with facial expression also depend on the chosen set of emotions, the nature of the database itself and the context of the interaction. Although the results obtained with facial expression only were lower than those obtained with speech only, this suggests that it is worthwhile to combine speech and facial expression, since there are emotions that are better recognized from the face than from speech [53]. In the case of the bimodal pairing consisting of speech and facial expression, the feature selection algorithm retained 22 features. When using facial expression and gesture, the performance of the classifier reached **65**% of instances correctly classified and the number of features remaining after feature selection was 15. Finally, the best pairing was that of the speech modality with the gesture modality. We obtained **75**% of instances correctly classified for this pairing, using 17 features. The best results that we obtained are in the case when all 3 modalities are fused. A somewhat surprising result is that using speech and facial expressions (62.5%) gives lower results than using facial expression and gesture (65%) or the best pairing, speech and gesture (75%). This suggests that, in our interaction scenario, the pairing of facial expressions and speech contains much less complementary information than the combination of facial expression and gesture, or of speech and gesture. Moreover, the result of the combination of facial expression and gesture (65%) is not an improvement over the results of gesture only (**67.1**%), whereas the combination of speech and facial expressions (62.5%) is an improvement over the results of speech only (57.1 %) and of facial expression only (48.3 %).

7 Discussion

As far as the results of the unimodal emotion recognition systems are concerned, the classifier based on body gesture data appears to be the most successful,

with **67.1**% of instances correctly classified. The overall performance of the classifier based on facial expressions information is **48.3**%, while the classifier based on speech data reaches **57.1**%.

The reason as to why the system trained with body gesture features proved to be the most successful may reside in the fact that, in the corpus of acted emotional expressions, each emotion is represented by a specific type of gesture: participants were provided with specific instructions in order to perform different gestures for each emotion. While this choice was made in order to build a system capable of recognizing different types of body gestures based on movement expressivity, it may have made the discrimination of emotions from body gesture easier than using facial and speech features.

Some of the results reported in the literature show higher recognition rates in systems which can infer emotions based on body gesture data. For example, Gunes and Piccardi [8] reported a recognition rate of 90% and Bernhardt and Robinson [10] built a system with an overall performance of 81%. While the results outperform those presented in this paper, it is worth noting that the current study aims to recognize 8 emotional states (contrary to the 6 inferred by the system in [8] and the 4 discriminated in [10]). Moreover, the unimodal system based on gesture presented in this study was trained and tested with a corpus of emotional expressions designed to be multimodal. This aspect, together with the fact that the emotional states were not expressed by professional actors, has probably influenced the way the body gestures were performed to express emotions, as well as the performance of the system.

In a similar manner to the system based on body gesture data, that based on facial features does not reach the recognition rates reported by some studies in the literature (see, for example, Littlewort et al. [54]). As discussed above, recognition rates should be interpreted in light of the characteristics of each specific system. Beside the differences determined by the number of emotions to be discriminated by the system, applying a unimodal classifier in a corpus designed for multimodal emotion recognition is likely to result in lower recognition rates compared to unimodal classifiers applied to corpus designed for unimodal analysis. This is particularly true for extreme acted expressions, which is the case for the work by Littlewort et al. [54], who used the Cohn and Kanade's DFAT-504 dataset. Moreover, the subjects in our dataset were not explicitly instructed about which facial expression to display while expressing emotions, a choice that resulted in a great variability of facial expressions for the same emotion, but which added to the naturalism of the corpus.

As far as the unimodal system based on speech data is concerned, in order to be able to compare the results of this study with some studies reported in the literature (see, for example, Schuller et al. [25], who reported a recognition rate of 70%), it is worth considering several aspects. One important aspect to consider is the selection process performed in order to obtain a corpus of speech units clearly assignable to the emotion classes to be discriminated. In [25] the authors report results on two well-known acted databases (Emo-DB and DES). While the authors obtain an average recognition rate of above 70% for the Emo-DB database, the results for the DES database reach only around 54%. There are many factors that could explain the difference between the two, and one of the most important is likely to be the unit selection process. While in the case of the emo-DB database, the selection has been done in order to use samples that were perceptually classified as more than 60% natural and at least 80% clearly assignable, the selection in the DES database produced samples that were reclassified in a perceptual test with an average accuracy of 67.3%. In this study, as no selection or perceptual evaluation has been performed, unfortunately there is no reference point to evaluate the effectiveness of the acted emotional expressions. A second aspect to consider, as discussed for the systems based on facial expressions and gesture data, is the number of emotions to be discriminated (8 emotion classes in this study versus 4 plus neutral in the DES database and 6 plus neutral in the Emo-DB database considered by [25]). Finally, the recording conditions play an important role as well, as the signal/noise ratio has a big influence on the quality of the features and, consequently, on the classification results. Although the recordings used in this study are not exceptionally noisy, they are not comparable to recordings made in studio conditions. As the scenario adopted in this study required the use of a microphone somewhat distant from the mouth, which was also not very directional, the signal-to-noise ratio cannot have been optimum.

The main objective of this study was to prove that using multiple modalities increases the performance of an emotion recognition system in the discrimination of 8 emotions. As expected, bimodal classifiers based on (1) both facial expressions and speech data and (2) both body gesture and speech data outperform the classifiers trained with a single modality. Nevertheless, the system based on facial expressions and body gesture data increases the performance of the system based on facial expressions only, but not the system based on body gesture data. This may be explained, as discussed above, by the fact that participants were not given instructions on which facial expression to display while expressing

emotions and this resulted in a higher variability of expressions for faces and, consequently, a significantly lower performance for the system based on facial features compared to that based on body gesture features. As we discussed for the unimodal systems, comparing the results of the bimodal systems with others reported in the literature is not an easy task. For example, Gunes and Piccardi [8] reported higher recognition rates for a system based on the integration of facial expressions and body gesture data. Yet, our results refer to a corpus of multimodal emotional expressions (three modalities vs. two in [8]) and a different scenario of interaction, in which more emotional states are taken into account.

As hypothesized, the fusing of all three modalities of data greatly improved the recognition rate in comparison with the unimodal and the bimodal systems: the multimodal approach based on feature level fusion showed an improvement of more than 10% compared to the performance of the system based on body gesture data and of more than 3% compared to the bimodal system based on body gesture and speech data. Further, the fusion performed at the feature level showed better performance than that performed at the decision-level, highlighting that processing modalities in a joint feature space is more successful, as expected by observing recent findings from psychology [55].

It is important to stress that, while all features used to train the classifiers describe how face, speech and body features vary over time, only some of them attempt to describe their dynamics. Consider, for example, the temporal profile of the Contraction Index of the body. While classical statistical features such as the maximum or the mean values only provide an overall description of the characteristics of this profile, other features such as the Initial Slope, the Final Slope and the Main Peak Duration divided by Duration convey information about dynamic aspects of the movement, such as the movement impulsivity. From the observation of the features conserved after the feature selection process (both in the unimodal and in the multimodal case), it appears that, independently of the modality of the expression, some of the features related to the dynamics of face, speech and body features are more effective than traditional statistical features in discriminating emotions (e.g., see the Initial Slope and the Main Peak Duration divided by Duration of the Contraction Index of the body, the Final Slope of the Velocity of the hand, the temporal range of the jaw opening, the timeline of the Bark spectral bands, etc.). This suggests that the dynamics of affective expressions is a crucial issue to be considered in emotion recognition.

The main contribution of this paper is to simultaneously use three modalities of expression for the recognition of 8 emotions. The performance of systems trained with all combinations of the three modalities is also tested for bimodal emotion recognition. To the best of our knowledge, no other study has addressed the issue of automated emotion recognition based on the face, body gesture and speech modalities in order to attempt to infer 8 emotions. Although several research works investigated the importance of gesture in emotion recognition systems (see, for example, [56], [10], [11], [12]) and a few studies have been successful in pairing of body gesture and facial expression for recognizing affective expressions (see, for example, Gunes and Piccardi [8], el Kaliouby and Robinson 2005 [36], Balomenos et al. [57]), the use of body gesture in this work is novel in the sense that no other study has added this modality when trying to improve the performance of an emotion detector based on facial expression and speech analysis.

Karpouzis et al. [58] [9] used data labelled using a continuous dimensional space (valence and activity) and mapped it onto 3 classes corresponding to 3 of the 4 quadrants in the valence-arousal plane (only 3 quadrants were necessary since little data was included in the left quadrant). This resulted in a 3-classes problem. One of the motivations to perform the current study was to investigate the issue of emotion recognition using a higher number of classes and, simultaneously, three modalities of expression. Karpouzis et al. [9] proposed a multi-cue approach based on facial expressions, speech and body gestures to infer affective expressions in naturalistic video sequences. In [9] the framework for the fusion of modalities includes facial expressions and speech data only, while the current study goes a step further by adding the body gesture modality so as to build a multimodal emotion recognition system.

Although it is true that, by their nature, several applications do not necessitate the use of body gesture, we did not aim to build a system intended to work in all contexts. We believe that body gesture can be applied to many applications where interaction plays a key role, such as virtual reality scenarios, interactive systems that can be used at home, in the office or in buildings, entertainment and artistic applications. In numerous interactive systems people need to gesture, since gesture is the only modality the system is based on. The interactive scenario described in this study is based on gestures that are part of a vocabulary that the user can exploit to interact with a machine which is able to analyze the way the user is asking for something (e.g., an object or a task to be accomplished by the machine itself). A system of this kind can be very useful in many applications. For these reasons, body gesture seems to be worthy of consideration as a useful modality in interactive systems based on emotion recognition.

8 Conclusion

This paper presented a multimodal framework for the analysis and recognition of emotions based on facial expression, body gesture and speech data.

The main contribution of this work is the integration of three modalities of expression for the recognition of emotions. In particular, the addition of body gesture information to facial expression and speech information for emotion recognition is novel. We also provided a thorough investigation, consisting of all combinations of two modalities, for the purpose of bimodal emotion recognition. As expected, results show that using three different modalities in combination greatly increases performance over unimodal emotion recognition systems. Further, the multimodal emotion recognition system is more effective than the systems trained with the combination of two modalities only. Humans use more than one modality to recognize emotions and process signals in a complementary manner, hence it was expected that an automatic system demonstrate similar behavior.

Consideration of multiple modalities is helpful when some modality feature values are missing or unreliable. This may occur, for example, when the feature detection process is made difficult by noisy environmental conditions, when the signals are corrupted during transmission, or, in an extreme case, when the system is unable to record one of the modalities. In real-life naturalistic scenarios, a system for emotion recognition must have the robustness to deal with these situations.

This work has highlighted the importance of the dynamics of emotional expressions. Facial expression, body movement and speech features included temporal features, i.e., features that retain information about the dynamics of the emotional expressions. As highlighted in the feature selection process, those features retaining dynamics information were, in numerous cases, selected as the most relevant for recognition purposes.

This study considered a restricted set of data, recorded in controlled conditions. Nevertheless, it represents a first attempt to fuse together three different synchronized modalities of expression, an approach often discussed, but still uncommon in current research. Considering a relatively small set of data before recording a larger set is very useful, since it allows for adjustment and modification of the data collection procedure in order to optimize it for the development of a larger corpus.

Future work will consider new multimodal recordings with a larger set of participants and, ideally, contain spontaneous expressions in real-life scenarios. In such scenarios, new challenges, including robustness to occlusions, noisy backgrounds (e.g., illumination changes, dynamic background), head motions, and so on, must be tackled more extensively.

Finally, an important issue to address in future work is the development of methods for multimodal fusion that take into account the mutual relationship between feature sets in different modalities, the correlation between audio-visual information and the amount of information that each modality conveys about the expressed emotion.

Acknowledgment

The research work has been conducted in the framework of the EU-IST Project HUMAINE (Human-Machine Interaction Network on Emotion), a Network of Excellence (NoE) in the EU 6th Framework Programme (2004-2007).

The authors would like to thank all the participants to the recording who kindly accepted to participate to the experiment.

We also thank Christopher Peters for proof-reading the english in the manuscript and providing additional constructive commentary.

References

1. M.F. Valstar, H. Gunes, and M. Pantic. How to Distinguish Posed from Spontaneous Smiles using Geometric Features. In *ACM International Conference on Multimodal Interfaces (ICMI'07), Nagoya, Japan, November 2007, Proceedings*, pages 38–45. ACM, 2007.
2. R. Picard. *Affective computing.* MIT Press, Boston, MA, 1997.
3. R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J.G. Taylor. Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, 20:569–571, January 2001.
4. N. Sebe, I. Cohen, and T.S. Huang. *Multimodal Emotion Recognition, Handbook of Pattern Recognition and Computer Vision.* World Scientific, ISBN 981-256-105-6, Boston, MA, 2005.
5. M. Pantic, N. Sebe, J. Cohn, and T.S. Huang. Affective multimodal human-computer interaction. *ACM Multimedia*, 20:669–676, November 2005.
6. N. Ambady and R. Rosenthal. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin*, 111(2):256–274, 1992.
7. C. Busso, Z. Deng, S. Yildirim, M. Bulut, C.M. Lee, A. Kazemzaeh, S. Lee, U. Neumann, and S Narayanan. Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proc. of ACM 6th int'l Conf. on Multimodal Interfaces (ICMI 2004)*, pages 205–211, State College, PA, October 2004.
8. H. Gunes and M. Piccardi. Bi-modal emotion recognition from expressive face and body gestures. *Journal of Network and Computer Applications*, 30:1334–1345, 2007.

9. K. Karpouzis, G. Caridakis, L. Kessous, N. Amir, A. Raouzaiou, L. Malatesta, and S. Kollias. Modeling naturalistic affective states via facial, vocal, and bodily expressions recognition. In *Artificial Intelligence for Human Computing*, 2007.
10. D. Bernhardt and P. Robinson. Detecting affect from non-stylised body motions. In A. Paiva, R. Prada, and R. W. Picard, editors, *Affective Computing and Intelligent Interaction, Second International Conference, ACII 2007, Lisbon, Portugal, September 12-14, 2007, Proceedings*, volume 4738 of *LNCS*, pages 59–70. Berlin: Springer-Verlag, 2007.
11. G. Castellano, S. D. Villalba, and A. Camurri. Recognising Human Emotions from Body Movement and Gesture Dynamics. In A. Paiva, R. Prada, and R. W. Picard, editors, *Affective Computing and Intelligent Interaction, Second International Conference, ACII 2007, Lisbon, Portugal, September 12-14, 2007, Proceedings*, volume 4738 of *LNCS*, pages 71–82. Berlin: Springer-Verlag, 2007.
12. A. Kleinsmith and N. Bianchi-Berthouze. Recognizing Affective Dimensions from Body Posture. In A. Paiva, R. Prada, and R. W. Picard, editors, *Affective Computing and Intelligent Interaction, Second International Conference, ACII 2007, Lisbon, Portugal, September 12-14, 2007, Proceedings*, volume 4738 of *LNCS*, pages 48–58. Berlin: Springer-Verlag, 2007.
13. L. Vidrascu and L. Devillers. Real-life Emotions Representation and Detection in Call Centers. In *in Proc. of 2nd International Conference on Affective Computing and Intelligent Interaction*, Lisbon, Portugal, 2005.
14. A. Batliner, S. Steidl, C. Hacker, E. Noth, and H. Niemann. Tales of tuning - prototyping for automatic classification of emotional user states. In *Proceedings of the Interspeech Conference*, 2005.
15. R. Banse and K.R. Scherer. Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, 70(3):614–636, 1996.
16. T. Vogt and E. Andre. Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition. In *Proc. IEEE International Conference on Multimedia and Expo ICME05*, 2005.
17. H. Gunes and M. Piccardi. A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior. In *Proc. of ICPR 2006 the 18th International Conference on Pattern Recognition*, Hong Kong, China, November 2006.
18. T. Banziger, H. Pirker, and K. Scherer. Gemep - Geneva multimodal emotion portrayals: A corpus for the study of multimodal emotional expressions. In *In L. Deviller et al. (Ed.), Proceedings of LREC'06 Workshop on Corpora for Research on Emotion and Affect, pages 15-019*, Genoa. Italy, 2006.
19. E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach. Emotional speech: towards a new generation of databases. *Speech Communication*, 40:33–60, 2003.
20. M. Rosenblum, Y. Yacoob, and L. Davis. Human expression recognition from motion using a radial basis function network architecture. *IEEE Transactions on Neural Networks*, 7(5):1121–1138, 1996.
21. M. Pantic and L.J.M. Rothkrantz. Automatic analysis of facial expressions: The state of the art. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(12):1424–1445, 2000.
22. M. Pantic and MS Bartlett. Machine analysis of facial expressions. In *Face Recognition*, K. Delac and M. Grgic, Eds., Vienna, Austria: I-Tech Education and Publishing, pp. 377-416, 2007.
23. R. Cowie and E. Douglas-Cowie. Automatic statistical analysis of the signal and prosodic signs of emotion in speech. In *In Proceedings International Conference on Spoken Language Processing*, Genoa. Italy, 1996.
24. P.Y. Oudeyer. The production and recognition of emotions in speech: features and algorithms. *International Journal of Human-Computer Studies*, 59(1-2):157–183, 2003.
25. B. Schuller, D. Seppi, A. Batliner, A. Maier, , and S. Steidl. Towards more reality in the recognition of emotional speech. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, pages 941—944, Honolulu, Hawaii, USA, 2007.
26. A. Camurri, I. Lagerlöf, and G. Volpe. Recognizing emotion from dance movement: Comparison of spectator recognition and automated techniques. *International Journal of Human-Computer Studies*, 59(1-2):213–225, July 2003.
27. N. Bianchi-Berthouze and A. Kleinsmith. A categorical approach to affective gesture recognition. *Connection Science*, 15(4):259–269, July 2003.
28. G. Castellano, M. Mortillaro, A. Camurri, G. Volpe, and K. Scherer. Automated analysis of body movement in emotionally expressive piano performances. *Music Perception*, 26(2):103–119, University of California Press, 2008.
29. R.W. Picard, E. Vyzas, and J. Healey. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23(10):1175–1191, 2001.
30. J. Kim, E. Andre, M. Rehm, T. Vogt, and J. Wagner. Integrating information from speech and physiological signals to achieve emotional sensitivity. In *Proc. of the 9th European Conference on Speech Communication and Technology*, 2005.
31. N. Sebe, I. Cohen, and T.S. Huang. Multimodal emotion recognition. *Handbook of Pattern Recognition and Computer Vision*, pages 981–256, 2005.
32. M. Pantic, N. Sebe, J.F. Cohn, and T. Huang. Affective multimodal human-computer interaction. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 669–676. ACM New York, NY, USA, 2005.
33. Z. Zeng, M. Pantic, G.I. Roisman, and T.S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):39–58, 2009.
34. Z. Zeng, J. Tu, M. Liu, T.S. Huang, B. Pianfetti, Roth D., and S. Levinson. Audio-visual affect recognition. *IEEE Transactions on Multimedia*, 9:424–428, 2007.
35. C. Busso and S. Narayanan. Interrelation between speech and facial gestures in emotional utterances: A single subject study. *IEEE Transactions on Audio, Speech, and Language Processing*, 20:2331–2347, November 2007.
36. R. el Kaliouby and P. Robinson. Generalization of a vision-based computational model of mind-reading. In *In Proceedings of First International Conference on Affective Computing and Intelligent Interfaces*, pages 582–589, 2005.
37. K.R. Scherer and H. Ellgring. Multimodal expression of emotion: Affect programs or componential appraisal patterns? *Emotion*, 7(1), 2007.
38. A.P. Engelbrecht, L. Fletcher, and I. Cloete. Variance analysis of sensitivity information for pruning multilayer feedforward neural networks. *Neural Networks, IJCNN '99*, 3:1829–1833, 1999.
39. D.J. Densley and P.J. Willis. *Emotional posturing: a method towards achieving emotional figure animation*, Computer Animation 1997.
40. M.-H. Yang, D. J. Kriegman, and N. Ahuja. Detecting faces in images: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(1):34–58, 2002.

41. J. W. Young. Head and face anthropometry of adult u.s. civilians. Technical Report final report, FAA Civil Aeromedical Institute, 1963-93.
42. S. Ioannou, G. Caridakis, K. Karpouzis, and S. Kollias. Robust feature detection for facial expression recognition. *EURASIP Journal on Image and Video Processing*, 2007.
43. A. Raouzaiou, N. Tsapatsoulis, K. Karpouzis, and S. Kollias. Parameterized facial expression synthesis based on mpeg-4. *EURASIP Journal on Applied Signal Processing*, 10:1021–1038, 2002.
44. A. Camurri, P. Coletta, A. Massari, B. Mazzarino, M. Peri, M. Ricchetti, Ricci, A., and G. Volpe. Toward real-time multimodal processing: Eyesweb 4.0. In *in Proc. AISB 2004 Convention: Motion, Emotion and Cognition*, Leeds, UK, March 2004.
45. A. Camurri, B. Mazzarino, and G. Volpe. Analysis of Expressive Gesture: The Eyesweb Expressive Gesture Processing Library, in A. Camurri, G.Volpe (Eds.), *Gesture-based Communication in Human-Computer Interaction, LNAI 2915*. Springer Verlag, 2004.
46. L. Kessous, N. Amir, and R. Cohen. Evaluation of perceptual time/frequency representations for automatic classification of expressive speech. In *International workshop on Paralinguistic Speech - between models and data, ParaLing'07*, 2007.
47. I.H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques, 2nd Edition*. Morgan Kaufmann, San Francisco, CA, 2005.
48. G. Cooper and E. Herskovits. A bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9(4):309–347, 1992.
49. I. Kononenko. On biases in estimating multi-valued attributes. In *In: 14th International Joint Conference on Articial Intelligence*, pages 1034–1040, Newcastle upon Tyne, UK, 1995.
50. R. Kohavi. A study on cross-validation and bootstrap for accuract estimation and model selection. In Morgan Kaufmann, editor, *Proceedings of the International Joint Conference on Articial Intelligence*, volume 2, pages 1137–1143, 1995.
51. L.S. Chen, T. S. Huang, Miyasato T., and Nakatsu R. Multimodal human emotion / expression recognition. In *Conf. on Automatic Face and Gesture Recognition*, 1998.
52. S. Ioannou, L. Kessous, and G. Caridakis. Adaptive on-line neural network retraining for real life multimodal emotion recognition. In *proceedings of International Conference on Artificial Neural Networks (ICANN)*, pages 81–92, Athens, Greece, September 2006.
53. L. C. De Silva, T. Miyasato, and R. Nakatsu. Facial emotion recognition using multimodal information. In *Conf. on Information, Communications and Signal Processing (ICICS'97)*, 1997.
54. G. Littlewort, M. Stewart Bartlett, I. R. Fasel, J. Susskind, and J. R. Movellan. Dynamics of facial expression extracted automatically from video. *Image Vision Computing*, 24(6):615–625, 2006.
55. B. Stein and M.A. Meredith. *The Merging of Senses*. MIT Press, Cambridge, USA, 1993.
56. M. Coulson. Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence. *Journal of Nonverbal Behavior*, 28(2):117–139, juin 2004.
57. T. Balomenos, A. Raouzaiou, S. Ioannou, A. Drosopoulos, K. Karpouzis, and S. Kollias. *Emotion Analysis in Man-Machine Interaction Systems*, pages 175–200. 3D Modeling and Animation: Synthesis and Analysis Techniques, Idea Group Publ., 2005.
58. K. Karpouzis, A. Raouzaiou, A. Drosopoulos, S. Ioannou, T. Balomenos, N. Tsapatsoulis, and S. Kollias. Facial expression and gesture analysis for emotionally-rich man-machine interaction. In *3D Modeling and Animation: Synthesis and Analysis Techniques*, 2004.